1 Introduction

The first step in answering a causal question is to be precise about the question you want to answer. In this first step, you will get to brainstorm a causal question you want to investigate. Picking a question that interests you will make the project a lot more enjoyable! It could even turn into a senior thesis eventually. We will need to analyze data to answer the causal question, so we will need to make sure that data on the topic is available. For that, you should look at the Add Health Data (described below) and think of a causal question involving the variables from that data set.

The next step in answering a causal question is clearly stating your assumptions about the causal system of interest. You'll also want to think about what factors you will need to adjust for to make conditional exchangeability plausible. You will get to draw a directed acyclic graph (DAG) that represents the causal system you proposed and consider what variables you need to condition on for a valid causal analysis. Make sure at least three relevant variables are available in the Add Health Data.

Fill in the answers to the "Action Items" section below using the provided .Rmd file. You are welcome to draw DAGs by hand instead of producing them by code. If you do this, you may scan or take a picture of your DAG and submit it as a separate file. Turn in Part 1 on Canvas by **Monday, Oct 20 at 11:59pm**.

2 Add Health

The National Longitudinal Study of Adolescent to Adult Health (often referred to as "Add Health") is a survey of roughly 20000 individuals which was first administered to high-school aged participants in 1994-1995. The survey followed up with the same participants at 4 subsequent time points, as recently as 2018 when the original participants were in their mid 20s or early 30s. The questions asked in the survey cover a range of topics including: Crime/Delinquency and Victimization, Demographic Characteristics, Education, Family, Medication and Substance Use and Abuse, Psychological Well-being and Cognition, Reproductive Health, SES, Labor Market and Occupation. You can find a more comprehesive list of topics and specific survey questions here. To get an idea of the types of questions people have used the data to answer, take a look at journal articles citing Add Health here.

3 Action Items

These are the questions you should answer and turn in for Part 1.

- (7 pts) Describe your causal question in a way that someone who has not taken this class would understand. Why are you interested in this question? How could answering this question allow for better decision making?
- (3 pts) What is the treatment? What is the outcome? Write out the potential outcomes using the notation we have used in class.
  - If your treatment is a variable that can take many different values, you could consider making it binary by simplifying the treatment in some way. For instance, if the treatment is the number of hours spent studying each day, you could dichotomize the treatment in the following way

$$A_i = \begin{cases} 0 & \text{if Hours } \leq 2 \\ 1 & \text{if Hours } > 2 \end{cases}$$

- (5 pts) How does the fundamental problem of causal inference apply to your question?

- (10 pts) We want to make sure there's some data available that gives you a chance to answer your question of interest. List out the variable names (as they appear on the ADD Health data) for the treatment and outcome below
  - Treatment variable name:
  - Outcome variable name:
- (5 points) List out 4-6 other variables that are particularly relevant for your causal question. These might be variables which either
  - Cause (either directly or through other factors) both your treatment variable and your potential outcomes
  - Are caused by both the treatment and the outcome

  Note, make sure at least 3 of the variables are included in the Add Health Data. List out the variable names as they appear in the ADD Health data.
- (10 points) Draw a DAG that includes your treatment variable, outcome variable, and the factors you listed in the previous part (including ones that might not be available in the data set). If you use letters for variables names, make sure you explain what the letter stands for.
- (5 points) In a few sentences, explain your DAG: tell us in words what is meant by each edge in your DAG.
- (5 points) Assuming your DAG is true, list one sufficient adjustment set to identify the causal effect of the treatment on the outcome. If a sufficient adjustment set does not exist, add additional variables to your DAG so that one does exist.